Supplementary Material for: The demographic histories of the M and S molecular forms of *Anopheles gambiae s.s.*

Jacob E. Crawford[1]* *and Brian P. Lazzaro*[1]

[1]*Department of Entomology, Cornell University, Ithaca, NY, USA.*

* *To whom correspondence should be sent: jc598@cornell.edu*

Supplementary Methods:

Molecular Form Datasets:

Our analysis is based on sequence polymorphism in coding fragments of 72 immune-related and 37 non-immune genes spread across all chromosome arms published by Cohuet et al. (2008). The mosquitoes sampled by Cohuet et al. (2008) are multiple *An. gambiae* individuals of the M (n=16 chromosomes) and S (n=18 chromosomes) molecular forms collected near Yaounde, Cameroon (03°51'N, 11°30'E). Mean nucleotide diversity was not significantly different between immune and non-immune loci and there was no evidence for strong selection in these data (Cohuet et al. 2008), so we consider all autosomal loci without regard to gene function in our analysis. We downloaded the heterozygous sequence fragments (accessions AM774672 – AM777160, AM900849 – AM900919), arbitrarily resolved the heterozygous sites to produce two hypothetical alleles for each individual and constructed alignments of each gene. Then, for each molecular form separately, the total number of segregating synonymous sites (S) was determined in each alignment and genetic diversity at synonymous sites was summarized for each molecular form as θ_W (Watterson 1975) and π (Tajima 1983) based on the total number of mutations using DnaSP (version 5.00.07, Librado and Rozas 2009). We used only synonymous sites to minimize any effects of natural selection in the dataset. θ_W and π are both estimators of the population parameter $4N_e\mu$ (Watterson 1975; Tajima 1983) where N_e is the effective

population size and μ is the neutral substitution rate, but they are calculated from different features of the empirical data. Whereas π is the average number of differences between alleles and is thus sensitive to allele frequency, θ_W is calculated based on the number of segregating mutations regardless of their frequency in the sample. π and θ respond differently to demographic shifts (Tajima 1989a,b). We used θ_W estimated from the empirical data to set the rate of mutation in the coalescent simulations, and we used π and S to summarize diversity in the simulated and empirical samples. These latter two summary statistics are the main components of the frequently used Tajima's D statistic (Tajima 1989a) and provide information about the shape of the underlying genealogy. Their relationship can be used to detect demographic (or selective) perturbations reflected in a sample. We used the components of the D statistic instead of the statistic itself because the D statistic is a biased summary of the data when recombination rates are not correctly incorporated and can compromise approximate-likelihood inference of demographic parameters (Thornton 2005). However, the bias is minimized if D is decomposed and its components, π and S, are used in its place (Thornton 2005). Only autosomal loci from the Cohuet et al. (2008) data set that were represented by at least ten alleles (range of 10 to 16 alleles for the M form and 10 to 18 alleles per locus) and exhibited a value of θ_W greater than zero were included our analysis (92 and 95 loci for M and S form respectively). Although excluding polymorphism-free loci from the analysis may slightly bias the dataset, a non-zero value of θ_W is needed for simulations (see below). We excluded X-linked loci for this analysis because large regions of the X-chromosome lack polymorphism, possibly due to recent selective sweeps (Stump et al. 2005; Turner et al. 2007), making most of the chromosome difficult to simulate.

Coalescent Simulations and Demographic Models:

We were interested in identifying a demographic scenario that can explain observed patterns of polymorphism in each of the molecular forms of *An. gambiae*. Our approach was to

simulate individual loci under specific population demographic scenarios, evaluate the fit of the simulated data to the empirical polymorphism data at individual loci, and combine these likelihood values into a 'genome likelihood.' We applied this approach to the M and S molecular forms independently. We modeled each gene individually by conducting 2×10^4 coalescent simulations under varying demographic scenarios using the program *ms* (Hudson 2002) conditioned on the sample size and θ_W for each gene as reflected in the empirical data set. Based on the fact that the sequenced genes are physically dispersed but typically shorter than 700 base pairs, we assumed free recombination between genes, but no intragenic recombination. Underestimating recombination is a conservative approach and not likely to produce large biases in the inference process. We calculated π and S from the *ms* output, which were then used to evaluate the 'genome likelihood' fit to the empirical data (described below).

We considered three families of demographic models: population growth, population bottleneck, and migration between two growing populations (main text fig. 1). For each model family, we explored a wide range of parameter values, chosen to be comprehensive but biologically plausible. The first model, population growth, varied in two parameters: the timing of the expansion (T_1, in units of $4N$ generations) and the ratio of ancient to current effective population size (N_{anc}/N_{curr}). The population bottleneck model family included the growth parameters listed above with the addition of a pre-expansion bottleneck that varied in both the severity of size reduction ($N_{pre\text{-}bottle}/N_{anc}$) and the number of generations the population remained at the reduced size (T_{bot}). The last model, migration between expanding subpopulations, included the growth parameters as well as migration from a second, unsampled subpopulation with growth parameters identical to the sampled subpopulation. The relative size of the unsampled subpopulation ($N_{unsampled}/N_{sampled}$) and rate of migration ($4Nm$) was also allowed vary in the model. The standard neutral drift-equilibrium model was considered as a null hypothesis. All parameter values are listed in Table 1.

Approximate Likelihood Method:

To determine how well each demographic model fit the empirical data, the simulated population samples were evaluated using an adaptation of Weiss and von Haeseler's (1998) approximate likelihood method. An indicator variable I_δ was calculated as

$$I_\delta(j) = \begin{cases} 1, & if, |\pi_{data} - \pi_{sim}| \le \delta_\pi \text{ and } |S_{data} - S_{sim}| \le \delta_S \\ 0, & otherwise. \end{cases} \quad (1)$$

where π_{data} equals the average number of pairwise differences in the empirical sample and S_{data} equals the number of segregating sites in the empirical sample at locus j. π_{sim} and S_{sim} were summary statistics calculated from the simulation results for that locus under the given model and δ_π and δ_S were positive numbers that define an empirically determined interval (see below) for locus j. Our method differs slightly from that of Weiss and von Haeseler (1998) in that they required the number of segregating sites to exactly match the empirical data, which is a slightly more conservative method, but we used the threshold approach to accommodate uncertainty in empirical estimates of the true population θ_W. The numerical threshold was designed to capture 20% of stochastic variation natural to the coalescent process, such that simulated values of π or S falling more than 10% above or below the empirical value resulted in the assignment of zero to the indicator variable (Weiss and von Haeseler 1998). Both threshold values were determined for each gene and each molecular form by conducting 2×10^4 coalescent simulations conditioned on the empirical sample size and θ_W for each locus under the standard neutral model. The summary statistics, π and S, were calculated for all simulations, assembled into a distribution and the thresholds were determined as the values 10% greater and less than the median of the simulated distribution.

The likelihood of the model given the data for each gene was estimated as the proportion of simulations that were assigned an I_δ of 1. The approximate likelihood function can be written as

$$lik(\Phi \,|\, \pi_{data}, S_{data}) \approx \frac{1}{B} \sum_{j=1}^{B} I_\delta(j), \quad (2)$$

where Φ is the model, π and S are summary statistics from the empirical data at locus j, B is the number of simulations (2×10^4) and I is the indicator variable from above. To obtain a genome-wide likelihood value that reflects the likelihood of the model given genome-wide patterns of polymorphism, gene-specific likelihoods were then natural log transformed and summed.

To identify the most likely model within each model family, we evaluated a series of models organized across a grid of parameter values using the above likelihood function to obtain the genome-likelihood value for each model. First, we searched a coarse grid of parameter values for each family of models. Next, in order to improve the precision of our parameter estimates, we adjusted parameter scales to finer levels in regions of the parameter space that showed high likelihood values in the coarse grid search and searched our finer-scale grid using the same likelihood procedure. We identified the best-fit model within each model family as the combination of parameter values that maximized the genome-wide likelihood function, and these best-fit models were then compared to determine which model family is most likely given the data (discussed below). To visualize the likelihood surface, we generated profile-likelihood curves for each parameter by plotting the maximum likelihood value for each parameter value. We estimated approximate 95% confidence intervals for each parameter using asymptotic theory where all parameter values with a likelihood value within 1.92 likelihood units (i.e. $X^2_{df=1}$ and $\alpha = 0.05$) of the maximum likelihood value were considered not significantly different from the MLE. Linear interpolation of the profile-likelihood curves was used where points were not simulated directly.

***Model Comparison*:**

After identifying the best-fit model within each model family, we compared models between families (e.g. growth vs. bottleneck) to identify the maximum likelihood estimate (MLE) for the demographic history of the *An. gambiae* population. We treated the models in a hierarchical fashion, with the standard-neutral model considered to be the primary null hypothesis. The simple growth model is an alternative to the standard neutral null. The more

complex bottleneck and migration models both have the growth model nested within them, so the growth model is considered the null model for testing bottleneck and migration hypotheses. Thus, the standard-neutral was first compared to the growth model. If the standard neutral null was rejected in this first comparison, the growth model then became the secondary null model against which the bottleneck and migration models were compared. If neither the bottleneck nor migration models fit significantly better than the growth model, we concluded that simple growth was the most parsimonious and likely model.

We compared models using the Akaike Information Criterion (AIC; Akaike 1974). Our models were not nested in the fashion required for evaluation of likelihood ratios. We employed AIC values to compare the likelihoods of non-nested models by penalizing models according to the number of free parameters in the model. We calculated AIC values as $AIC_i = -2(lmax_i - k_i)$ where $lmax_i$ is the maximum likelihood value under model i and k is the number of free parameters in model i, such that a higher AIC value means a better fit to the data (Akaike 1974). Then we used the statistic $\Lambda = AIC_{alt} - AIC_{null}$ to compare AIC values between models (Caicedo et al. 2007). Negative values of this statistic indicate that the alternative model is a better fit. We established a null distribution by simulating 10^4 'genomes' comprised of the same number of loci as the empirical dataset under the null model, evaluating the maximum likelihood of each 'genome' under the null and alternative models and calculating Λ_{sim} as the difference between the AIC statistics calculated under the null and alternative models. We calculated a p-value as the proportion of simulations with $\Lambda_{sim} < \Lambda_{obs}$.

***Model performance*:**

Although the approximate likelihood method used here explicitly evaluates the fit of the model to the entire dataset, we wanted to confirm that our best-fit model is able to adequately reproduce the empirical data for each molecular form. To this end, we simulated all chromosome III loci under the best-fit migration model for each molecular form and plotted the median value of π

and S from 10^4 coalescent simulations next to the empirical data (Supplementary figs. 4 and 5). Simulations were conducted as above where each locus was simulated using *ms* conditioned on the empirical sample size and θ_W. The distributions of the summary statistics are often skewed so we compared the median value to the data in order to minimize biases associated with mean values of skewed distributions. We considered the comparison of loci on only one chromosome sufficient to demonstrate the adequacy of model performance and arbitrarily chose chromosome III.

***Comparison of the timing of expansion between Molecular forms*:**

To determine whether the MLE timings of expansion were significantly different between the molecular forms, we asked whether the timing inferred for one molecular form was within the confidence region of the timing or growth parameter (T_1) estimated for the other molecular form. For example, the inferred timing of growth for the M form is $3.0N_{curr}$, which corresponds to $2.1N_{curr}$ for the S form after calibration for the relative effective population sizes (we estimate that the M form is 0.7 times the S form). This value is outside of the confidence interval estimated for the timing of growth for the S form (95% C.I. 2.18 - 2.88), suggesting that this more recent timing of the M form expansion did not overlap in time with the S form expansion.

***Population genetic re-analysis of Obbard et al. (2009) data*:**

To test the effects of applying the demographic correction to the null model on population genetic analyses, we compared Tajima's D values from 4 serpin loci and 4 control loci obtained by Obbard et al. (2009) first to null distributions simulated under the standard-neutral equilibrium (SNE) model then to null distributions simulated under the MLE migration models inferred here. We simulated each sample and locus individually using the same simulation framework described above. 10^4 coalescent simulations were conducted using the

coalescent simulation program *ms* (Hudson 2002) for each locus-population combination conditioned on the number of chromosomes sampled for that locus-population combination and θ_W estimated from the empirical data. Tajima's *D* was calculated from each simulated sample and assembled into null distributions for a given locus-population combination. Null distributions were generated both under the standard-neutral equilibrium as well as under the MLE migration models for each form. Empirical *D* values were then compared to null distributions in a one-tailed test, the polarity of which depended on whether *D* was positive or negative. No correction for multiple testing was made, so *D* was considered significantly unlikely under a given null model if the empirically observed value fell into the 5% tail of the null distribution. The simulations assumed no recombination, which is a reasonable approximation given the short sequences (range of 354 to 783 basepairs), and so are conservative with regard to testing hypotheses of selection.

Simulations of N_e-adjusted migration models:

One possible explanation for the better fit of migration models is that the effective population size is increased through migration, and thus no migration is actually necessary in the models. To determine whether manually adjusting the effective population size can account for the increased likelihood of the migration models over the growth models, we simulated each locus under the MLE growth model, but we adjusted θ_W to reflect the larger effective population size. For example, the MLE migration model for the M molecular form includes migration between the sampled population and an unsampled population that is 0.4 times the size of the sampled population, so we multiplied the empirical θ_W for each locus by 1.4 so that the adjusted simulated population is one panmictic unit 1.4 times as large as its unadjusted counterpart. These adjusted models were compared to the unadjusted MLE growth and MLE migration models using the model comparison framework described above.

Literature Cited:

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716-723.

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, et al. 2007. Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. PLoS Genetics 3: e163.

Cohuet A, Krishnakumar S, Simard F, Morlais I, et al. 2008. SNP discovery and molecular evolution in *Anopheles gambiae*, with special emphasis on innate immune system. BMC genomics 9: 227.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

Obbard DJ, Welch JJ, Little TJ. 2009. Inferring selection in the Anopheles gambiae species complex: an example from immune-related serine protease inhibitors. Malaria Journal 8: 117.

Stump AD, Fitzpatrick MC, Lobo NF, Traoré S, et al. 2005. Centromere-proximal differentiation and speciation in Anopheles gambiae. Proc. Natl. Acad. Sci. 102: 15930–15935.

Tajima F. 1989a. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Tajima F. 1989b. The effect of change in population size on DNA polymorphism. Genetics 123: 597-601.

Thornton K. 2005. Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. Genetics 171: 2143-2148.

Turner TL, Hahn MW. 2007. Locus- and population-specific selection and differentiation

between incipient species of Anopheles gambiae. Molecular Biology and Evolution 24: 2132-2138.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7: 256-276.

Supplementary Table 1: Demographic model parameter ranges, sampling density and rejection statistics

Parameter	Range[a] (units)	Growth		Bottleneck		Migration	
		M	S	M	S	M	S
Fold population expansion (N_{anc}/N_{curr})	0 – 10,000	42	33	22	19	11	11
Generations since growth (T_1)	0.05 – 1.2 ($4N_{curr}$ generations)	64	55	29	29	17	13
Fold population reduction during bottleneck ($N_{pre\text{-}bottle}/N_{anc}$)	1.25 – 10,000	---	---	4	4	---	---
Duration of bottleneck (T_{bot})	0.01 – 0.5 ($4N_{curr}$ generations)	---	---	4	5	---	---
Subpopulation size ($N_{unsampled}/N_{sampled}$)	0.1 – 1.0	---	---	---	---	9	8
Rate of migration (*4Nm*)	10^{-4} – 10 (migrants per generation)	---	---	---	---	10	10
Total number parameter combinations[b]		2,688	1,815	7,888	7,556	16,830	11,440
Percentage of simulations accepted[c]		9.74	10.66	9.42	10.17	8.28	9.47
Total number parameter combinations accepted[d]		432	267	1,269	1,062	88	149
Percentage of simulations accepted within accepted models[e]		10.39	11.64	10.40	11.66	11.29	13.63

[a] For each demographic model and molecular form, we searched the parameter space uniformly over coarse intervals. We then adjusted the parameter space to include a higher density of grid points (parameter combinations) in the region with the highest likelihood values in the first search and evaluated the grid a second time.

[b] Total number of parameter combinations searched in grid after increasing density of parameter values sampled in the second grid search.

[c] Percentage of all simulations that was not rejected within likelihood framework. Each locus was simulated 20,000 times for each parameter combination.

[d] Total number of parameter combinations that received likelihood value within 1.92 likelihood units of the maximum.

[e] Percentage of simulations within accepted models (see [d]) that were not rejected within the likelihood framework.

Supplementary Table 2: Population genetic re-analysis of Obbard et al. (2009) data under SNE and MLE migration models

		Tajima's *D*		
Locus	Population[a]	*D*	*SNE*[b]	*MLE*[c]
Control 4	Burkina Faso	0.16	0.3715	0.1894
	Kenya	0.34	0.3066	0.0622
Serpin 4C	Burkina Faso	-0.11	0.5093	0.6333
	Kenya	0.90	0.1575	**0.0372**
Control 5	Burkina Faso	-1.74	**0.0235**	**0.0160**
	Kenya	0.36	0.3123	0.1784
Serpin 5	Burkina Faso	-1.33	0.0737	0.0609
	Kenya	0.07	0.4158	0.2077
Control 6	Burkina Faso	-1.85	**0.0113**	**0.0076**
	Kenya	---	---	---
Serpin 6	Burkina Faso	-0.63	0.2943	0.3361
	Kenya	1.25	0.0836	**0.0069**
Control 16	Burkina Faso	-0.75	0.2569	0.2850
	Kenya	---	---	---
Serpin 16	Burkina Faso	-0.29	0.4284	0.5539
	Kenya	-0.77	0.2507	0.3504

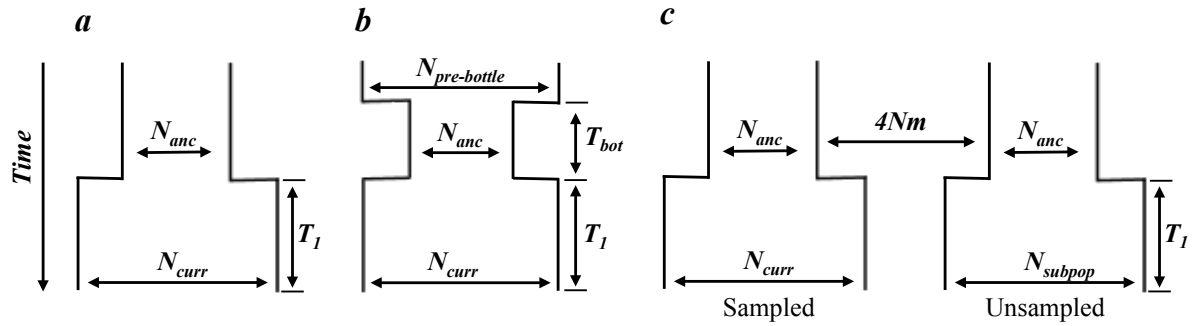
[a] location where *An. gambiae* were sampled

[b] *P* values indicating probability of statistic when compared to null distribution simulated under standard-neutral equilibrium

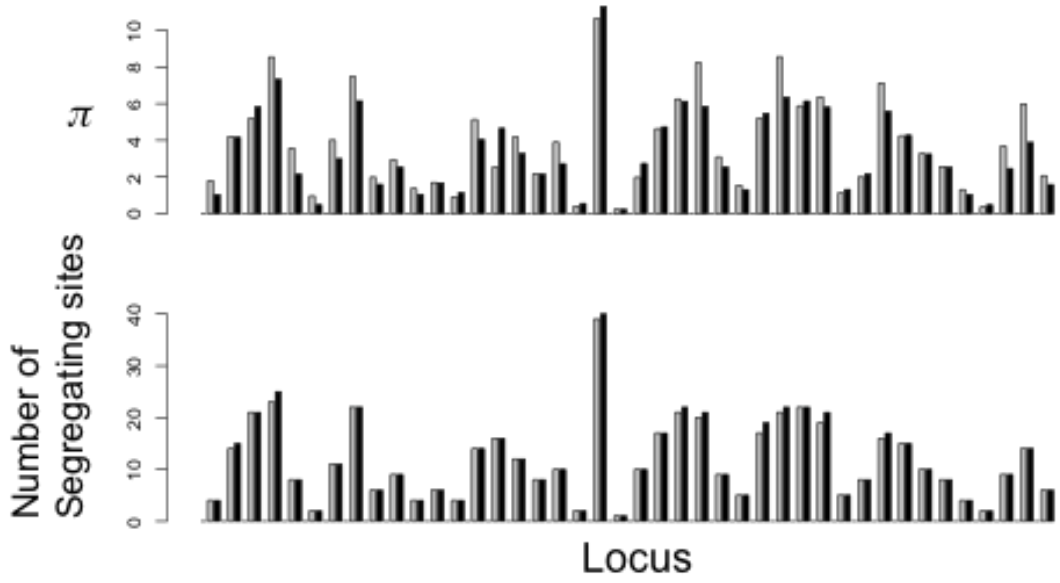
[c] *P* values indicating probability of statistic when compared to null distribution simulated under MLE migration model

- Values in bold font were significantly inconsistent with the simulated null model at the nominal 5% threshold (no correction for multiple testing)

Supplementary Figure S1: Demographic models and their varied parameters *(a)* Population growth included time of expansion (T_1) and size of expansion (N_{curr}/N_{anc}) variables *(b)* Population bottleneck included growth parameters (T_1 and N_{curr}/N_{anc}), the size of population reduction and duration of bottleneck (T_{bot}) *(c)* Migration between growing subpopulations including growth parameters (T_1 and N_{curr}/N_{anc}), the rate of migration ($4Nm$) and the size of the unsampled subpopulation relative to the sampled subpopulation.



Supplementary Figure S2: Comparison of M form loci modeled under the MLE to M form empirical data. We conducted 10^4 simulations using the program *ms* under the MLE migration model for each 3rd chromosome locus and plotted the median value (gray bars) of the average number of pairwise differences (π) and the number of segregating sites (S) next the empirical value (black bars) of each statistic for that locus. No intralocus recombination was included in the simulations. Loci are ordered according to their relative positions on the 3rd chromosome.



Supplementary Figure S3: Comparison of S form loci modeled under the MLE to S form empirical data. We conducted 10^4 simulations using the program *ms* under the MLE migration model for each 3rd chromosome locus and plotted the median value (gray bars) of the average number of pairwise differences (π) and the number of segregating sites (S) next the empirical value (black bars) of each statistic for that locus. No intralocus recombination was included in the simulations. Loci are ordered according to their relative positions on the 3rd chromosome.